**Title: Development and Evaluation of the British English Coordinate Response Measure Speech-In-Noise Test as an occupational hearing assessment tool**

Hannah D. Semeraro1, Daniel Rowan1, Rachel M. van Besouw1, Adrian A. Allsopp2

1Institute of Sound and Vibration Research, University of Southampton, UK.

2Institute of Naval Medicine, Gosport, UK.

**Key words**: Speech Perception; Behavioural Measures; Psychoacoustics/ Hearing Science; Noise; (Auditory Fitness For Duty)

**Acronyms**: auditory fitness for duty, AFFD; coordinate response measure, CRM; decibels, dB; graphical user interface, GUI; mission-critical auditory task, MCAT; North Atlantic Treaty Organisation, NATO; noise induced hearing loss, NIHL; psychometric functions, PF; pure tone audiometry, PTA; speech-in-noise, SiN; signal to noise ratio, SNR; speech reception threshold, SRT; triple-digit test, TDT.

**Corresponding Author**

*Hannah Semeraro, the original corresponding author, is on maternity leave until September 2017 so please direct any further correspondance to co-author Daniel Rowan, continuing to cc Hannah Semeraro.*

Daniel Rowan & Hannah D. Semeraro

dr@isvr.soton.ac.uk & h.d.semeraro@soton.ac.uk

Institute of Sound and Vibration Research

University of Southampton

United Kingdom

SO17 1BJ

# Abstract

*Objective:* The studies described in this paper outline the design and development of a British English version of the coordinate response measure (CRM) speech-in-noise (SiN) test. Our interest in the CRM is as a SiN test with high face validity for occupational auditory fitness for duty (AFFD) assessment.

*Design:* Study 1 used the method of constant stimuli to measure and adjusted the psychometric functions of each target word, producing a speech corpus with equal intelligibility. After ensuring all the target words had similar intelligibility, for studies 2 and 3 the CRM was presented in an adaptive procedure in stationary speech-spectrum noise to measure speech reception thresholds and evaluate the test-retest reliability of the CRM SiN test.

*Study sample*: Studies 1 (n=20) and 2 (n=30) were completed by normal-hearing civilians. Study 3 (n=22) was completed by hearing impaired military personnel.

*Results:* The results display good test-retest reliability (95% confidence interval < 2.1 dB) and concurrent validity when compared to the triple-digit test (*r = < 0.65)* and the CRM is sensitive to hearing impairment.

*Conclusion:* The British English CRM using stationary speech-spectrum noise is a ‘ready to use’ SiN test, suitable for investigation as an AFFD assessment tool for military personnel.

# 1. Introduction

The accurate assessment of occupational fitness for duty is of utmost importance across a range of jobs to ensure an individual is able to carry out their work effectively and without posing a risk to their own or others’ health and safety. The capability to maintain situational awareness is a vital aspect of fitness for duty for a number of occupations. Situational awareness is a term used to describe a person being aware of what is happening around them and able to calculate the relative importance of difference aspects of their surroundings (Endsley, 1988). There are three main components of situational awareness: 1) information gathering, 2) perception of the information and 3) making sense and giving meaning to these perceptions. Loss of situational awareness can be caused by problems at any one of these stages, but of particular interest here is encountering difficulty during the information gathering stage due to hearing loss.

A large number of occupations require hearing assessment tools that are able to accurately assess auditory fitness for duty (AFFD), defined by Tufts et al (2009) as the possession of sufficient hearing abilities for safe and effective job performance. Police officers, fire fighters, military personnel and factory workers are all examples of employees who are potentially exposed to damaging levels of noise as part of their work and they are also at risk of significant consequence to health and safety if their hearing impairment affects their ability to do their job safely and effectively. Currently, pure-tone audiometry (PTA) is the standard tool used to measure hearing in an occupational context. In terms of assessing AFFD, there are three main reasons to suggest that PTA is not a suitable tool for achieving this and these are discussed further below.

## *A measure of AFFD should assess hearing ability not acuity*

PTA measures tone detection in quiet, but this assessment method only measures the audibility aspect of hearing impairment and does not assess the distortion component associated with sensorineural hearing loss (Plomp, 1978). Measures of AFFD should predict performance on hearing critical tasks, which a measure of audibility is not achieving. For example, in the military context, a previous study identified operationally important tasks for which good hearing was considered to be essential by infantry personnel returning from the war in Afghanistan. These tasks are referred to as mission-critical auditory tasks (MCATs) (Bevis et al, 2014; Semeraro et al, 2015). The majority of the MCATs (seven out of nine) involve speech communication in adverse conditions, typically involving background noise. This suggests that a measure of AFFD needs to be able to accurately assess speech in noise (SiN) performance.

PTA has been shown to be a poor predictor of performance for speech intelligibility in noise (e.g. Smoorenburg, 1992; Peters et al, 1998; Leensen, 2013) and is likely to be so for other adverse listening conditions such as job specific hearing critical tasks. The difficulty PTA has in predicting speech intelligibility in adverse conditions for those with sensorineural hearing loss is partly because sensorineural hearing loss is associated with a variety of supra-threshold deficits (Akeroyd, 2008; Füllgrabe et al, 2014; Shinn-Cunningham et al, 2014). In addition, the topic of ‘hidden hearing loss’ is currently prominent in auditory research, referring to measurable deficits on speech intelligibility tests but pure tone thresholds within normal clinical limits (Plack et al, 2014). There is evidence from animal studies to suggest that individuals exposed to high sound levels (such as the military population) can have reversible threshold shifts but permanent cochlear synaptopathy, which is likely to cause performance deficits when listening in complex situations, such as speech intelligibility in noise (Kujawa & Liberman, 2015).

Although the impact of supra-threshold deficits on performance for speech communication tasks is still not fully understood, a SiN test is more likely to capture how these deficits affect performance on speech communication tasks and thus provide a more accurate estimation of AFFD in comparison to PTA. Similar logic has been followed by others interested in detecting functional hearing loss in the military. Tufts et al (2009) recognised that AFFD protocols should include measures beyond PTA, Brungart and Sheffield (2013) explored the use of the Modified Rhyme Test for predicting performance on operational tasks, Giguère et al (2008) developed the Canadian-French Hearing in Noise Test as an AFFD screening tool for hearing-critical jobs and Leensen et al (2011a; 2011b) have developed a version of the triple-digit test as a screening tool for detecting high-frequency hearing loss, as observed in noise exposed individuals.

## *Measures of AFFD should hold high face validity*

Face validity can be described as “whether a test ‘looks valid’ to examiners who take it, the administrative personnel who decide on its use and other technically untrained observers” (Anastasi, 1988, p.144). For certain types of measurement face validity can be of particular importance and one such area of testing is diagnostic tools. The World Health Organisation list ‘acceptability’ as a key factor to take into account for any health screening test (WHO, 2016). This increases the likelihood of test acceptance by non-specialists as an appropriate measurement tool, increasing the potential for implementation. This makes it easier for the clinician to explain test results and how they link to performance in an operational environment, allowing personnel to begin to understand and accept the impact of their hearing loss. In the context of AFFD assessment, listening to quiet beeps, as with PTA, does not clearly reflect the skills required to carry out many job specific hearing critical tasks and therefore an individual’s ability to do their job safely and effectively. Introducing a tool that is clearly assessing the skills required to carry out a job may increase the likely uptake of the test across a range of occupations and acceptance of the test by employers and employees.

## *Practicalities of running AFFD assessment*

PTA requires a specially trained operator who can accurately conduct audiometry and correctly interpret the results. In addition, specialist equipment and a sound proof environment are required so normally only one employee can be assessed at a time. SiN testing can be carried out without the need for a soundproof environment and testing can be fully automated, without the need for a specially trained operator. SiN tests do not measure absolute threshold so the ambient noise level only needs to be below the noise level of the test stimuli to ensure it does not interfere with the results. SiN tests can be run through a computer tablet application, removing the need for specialist equipment and allowing for multiple employees to be tested at once, reducing testing time. Finally, the resulting speech recognition threshold (SRT) is only a single number for left and right ears that is easy to interpret to decide what further action is needed.

The above considerations are applicable to all occupations in which assessing performance on hearing critical tasks in order to regularly and accurately measure AFFD is important. This paper now focusses specifically on developing an AFFD assessment tool for the UK Armed Forces, addressing the above considerations and providing a framework for SiN test selection, development and validation for a range of other occupations.

# 2. Selection and recording of test material for a military-specific AFFD SiN test

When considering which SiN test to further explore as a potential AFFD assessment method for the UK military there are several factors to consider. Firstly, from a practical viewpoint, selecting closed-set test stimuli means the test can be self-scoring, reducing the need for a specially trained operator and increasing the potential for multiple tests to be conducted simultaneously. This could lead to time and cost saving implications. Secondly, choosing a test with high face-validity will contribute towards test acceptance and help personnel to understand how their results link to job performance. For this reason, selecting speech stimuli that are similar to military command structure is important.

A review was carried out to explore the similarity between speech tests and military communications. Using opportunistic sampling six military personnel were asked to rank, in their opinion, how relevant different speech test stimuli were in comparison to military communication. These personnel had been in active service and had knowledge of command structure from their basic infantry training. The survey included seven examples of SiN tests, selected to provide an overview of different types of speech stimuli (the Bamford-Kowal-Bench (BKB) sentence test, Bench et al, 1979; the Coordinate Response Measure (CRM), Kitterick et al, 2010; the Hearing in Noise Test, Nilsson et al, 1994; the Quick Speech in Noise (QuickSIN) test, Taylor, 2003; the Triple Digit Test (TDT), Lutman et al, 2006; and the Words In Noise (WiN) test, Wilson and Cates, 2008). They were given Likert scale ratings: 1) very relevant to military communications; 2) has some relevance to military communications; 3) no relevance to military communications. The median scores for each were calculated and the WiN, TDT and CRM stimuli were all, on average, ranked as being ‘very relevant to military communication’. This indicates, albeit from a small sample and an un-validated survey, that these stimuli hold high face validity to military communication. Of these three the CRM was chosen for further investigation due to the target words reflecting military communication and it being simple to implement in an automated, self-scoring adaptive procedure.

The CRM is a SiN test that requires listeners to detect target words (a call sign, colour and number) within the carrier phrase ‘Ready *call sign,* go to *colour number* now’(Bolia et al, 2000). It has been previously used for studies of speech intelligibility and sound-source localisation, usually with the call sign as an indicator of which sentence to listen to when multiple sentences are presented simultaneously (e.g. Bolia et al, 2000; Brungart, 2001a; Brungart, 2001b; Eddins & Liu, 2012). Our interest in the CRM is as a SiN test with high face validity for occupational AFFD assessment, with a particular focus on the military, but also including other workforces which use ‘command and control’ instructions, such as the police, firefighters and coastguards.

There are currently two versions of the CRM: one with an American English speaker (Bolia et al, 2000) and one with a British English speaker (Kitterick et al, 2010). There are two key reasons why a decision was made that a new version of the CRM was required in order to meet the requirements of a tool to be used to assess AFFD within the British Ministry of Defence.

Firstly, prior to using speech stimuli within an adaptive procedure it is important that some key assumptions are met and according to Levitt (1971) the most important of these is that there is a monotonic relationship between SNR and performance. In order for this monotonic relationship to be observed it is necessary for all the stimuli to be homogenous, i.e. there is stability between performance levels across SNRs between target words (Leek, 2001). This is achieved by measuring and adjusting the intelligibility of each target word to ensure equal speech intelligibility across the speech corpus and is a recognised process when developing SiN tests (e.g. the McCormick Automated Toy Test, Summerfield et al, 1994; the Triple Digit Test, Smits et al, 2004 & Ozimek et al, 2009). For both the British (Kitterick et al, 2010) and American English (Bolia et al, 2000; Eddins & Liu, 2012) versions of the CRM there are currently no published data on the equalisation of the target words in terms of intelligibility, the test-retest reliability of the stimuli in an adaptive procedure and sensitivity to hearing impairment. For both of these versions the RMS amplitude of the stimuli have been equalised and a monotonic relationship has been assumed prior to using the stimuli in an adaptive procedure. In this study it was considered important to obtain speech intelligibility measures for the individual target words and to adjust the amplitude of the stimuli in order to equalise the intelligibility of the CRM test material.

Secondly, neither of the current versions of the CRM use the North Atlantic Treaty Organisation (NATO) phonetic alphabet that is used by the British Armed Forces, emergency services, air traffic control and radio operators; for example, they use *Arrow, Eagle, and Tiger* instead of the NATO alphabet *Alpha, Echo and Tango*. In terms of selecting a SiN test with high face validity it was considered important to ensure the language used within the test accurately represented British military communication. Considering that the process of intelligibility equalisation was required regardless of which speech corpus was used it was decided that the additional step of recording the CRM using the NATO phonetic alphabet was worthwhile in the development of a SiN test with high face validity to be used by the UK Armed Forces. This paper describes the necessary stages of preparing these stimuli to be used in an adaptive procedure.

We developed a British English version of the CRM using 18 disyllabic call signs from the NATO phonetic alphabet, nine monosyllabic numbers and nine monosyllabic colours. The new speech materials were recorded and assessed for quality and were adjusted in level to ensure equal intelligibility in noise (Section 4: Study 1). An adaptive version of the CRM with speech-spectrum noise was assessed for test-retest reliability and concurrent validity with normal-hearing civilian (Section 5: Study 2) and hearing-impaired military (Section 6: Study 3) participants. The results of Study 3 also provide an initial indication of the sensitivity of the test to military-related hearing impairment as characterised by PTA.

The British English CRM speech corpus is outlined in Table 1. The recordings were made in an acoustically treated chamber at the Institute of Sound & Vibration Research and in three blocks (“Ready *call sign*”, “go to *colour*” and “*number* now”) so they could be concatenated to create all possible sentences. The speaker was the second author, a male with a standard southern English accent similar to Received Pronunciation (The British Library, 2015). Mono 16-bit 44100 Hz sound files were recorded with a RME BabyFace soundcard and Brüel & Kjær type 2230 sound level meter positioned 0.75 m away from the speaker at head height. A minimum of six recordings of each target word were made. One version of each was then selected based on the authors’ subjective impression of clarity, intonation, loudness and naturalness when concatenated to form a sentence. Each recording was adjusted to the same RMS amplitude; these recordings are referred to as the ‘RMS amplitude adjusted recordings’.

A 28 s stationary speech-spectrum noise was generated in MATLAB using Gaussian white noise filtered to have the same long-term average spectral shape as the CRM sentences. A random, independent segment from this was selected when presented together with a CRM sentence. This noise type was chosen because it has fixed amplitude which is beneficial when not testing in a sound proof environment, reducing the likelihood of ambient noise interfering with the test as may be the case with arguably more realistic but intermittent maskers, such as babble noise.

# 3. General methods

Participants were seated in a soundproof booth (Study 1 and 2) or quiet classroom (Study 3) and were presented with CRM sentences in stationary speech-spectrum noise bilaterally via Sennheiser HDA 300 headphones. The CRM sentences were concatenated, presented and scored using custom MATLAB code with a graphical user interface (GUI). Participants indicated the target words they heard using the GUI. For all studies a forced choice method was used; participants were required to give an answer before proceeding to the next sentence. Before testing, participants were presented with five sentences at an advantageous SNR to familiarise themselves with the GUI and stimuli.

The noise level was fixed at 63 dB A for Studies 1 and 2, which involved normal-hearing participants. For Study 3, which involved hearing-impaired listeners the noise level was adjusted by the individual listener to ensure the overall presentation level was supra-threshold (further details in Section 6). For all studies throughout testing the signal to noise ratio (dB SNR) determined by varying the level of the speech and the noise level was kept constant.

The hearing of participants is described as the average PTA thresholds of their better hearing ear, taken across 0.25, 0.5, 1, 2, 3, 4, 6 and 8 kHz. ‘Normal-hearing’ is defined as PTA thresholds <20 dB HL across all frequencies and both ears. All studies were approved by the University of Southampton Ethics Committee and Research Governance Office and Study 3 was additionally approved by the Ministry of Defence Ethical Committee.

# 4. Study 1: measuring and equalising CRM speech intelligibility functions

Some words are more intelligible than others when presented at the same SNR. This is undesirable for a test like the CRM where it is necessary for a listener’s score to be independent of the target words presented during the test. The aim of Study 1 was to ensure that all target words had similar intelligibility when presented at the same SNR. To achieve this, psychometric functions (PFs) were measured for each of the target words in noise using the RMS amplitude adjusted recordings. The levels of the words were further adjusted where necessary in an attempt to produce PFs with the same average speech reception threshold, defined as the SNR at which participants achieved 50% correct (SRT50). The PFs were then re‑measured to check that this had been accomplished. This process also allows for any words whose PF differs markedly from the others in terms of unadjusted SRT50 or slope to be rejected.

# 4.1. Study 1 methods

Twenty normal-hearing participants (10 male, 10 female, mean age 26 years) attended three sessions. Two participants withdrew between sessions two and three. The experimental procedure in all three sessions was identical and used the method of constant stimuli. Each listener received 54 sentences at each SNR. Across those 54 sentences, each call sign was presented three times, and each colour and number was presented nine times. For this study only, responses for each individual target word within a sentence were recorded separately. This was carried out at seven SNRs (–1, –4, –6, –9, –12, –14 and –17 dB), selected following a pilot study with six different participants. The order and formation of concatenated sentences at each SNR was randomised.

Sessions 1 and 2 were identical and used the RMS amplitude adjusted recordings. The results were pooled across both sessions and all participants, resulting in a percentage correct score for each target word at each SNR. A logistic PF was fitted to the data with the least squares method using Kingdom and Prins’s (2010) MATLAB toolbox. The SRT50 was found from the PF for each target word. If there was a difference of ±1.5 dB or more between a target word SRT50 and the mean SRT50 for the corresponding target word group (call sign, colour, number) then the level of the target word was adjusted to within 0.1 dB of the target word group mean SRT50. Session 3 was identical to Sessions 1 and 2 except that it used the SRT50 matched recordings.

# 4.2 Study 1 results and discussion

Figure 1 plots the PF for each individual word before (left column) and after (right column) level adjustment. The pooled results of Sessions 1 and 2 have a higher standard deviation (SD) of SRT50 across the words within each word group in comparison to the Session 3 results (SDs for Sessions 1 and 2 (1,2) compared to Session 3 (3): call signs = 1,22.7, 30.8 dB; colours = 1,21.7, 30.9 dB; numbers = 1,22.5, 30.6 dB). The bootstrap analysis available in the Kingdom and Prins’s (2010) MATLAB toolbox was used to determine the 95% confidence interval (CI) of the SRT50 for individual words for Session 3. The 95% CI for each target word was compared to corresponding target word group average SRT50 (calculated as the mean SRT50 score for all target words within a word group). All the target word 95% CIs were found to be within ±1.5 dB of the corresponding word group. A visual inspection of the PF slopes in Figure 1 shows no concerning discrepancies between the shape of the slopes within each target group or across the three sessions. The average SRT50 for each word group differed by less than 0.6 dB between Sessions 1 and 2, and Pearson correlation coefficients comparing SRT50 for individual words between those two sessions were higher than 0.95. In summary, the PFs with the SRT50 matched recordings were considered sufficiently similar to enable the speech material to be utilised in an adaptive procedure, the success of which was assessed in Study 2.

# 5. Study 2: evaluation of the adaptive CRM with normal-hearing listeners

The CRM speech material were presented in an adaptive procedure for estimating an individual’s SRT in a manner that is automated, quick and accurate (Leek, 2001). Depending on the use of the test, one might or might not want to score the call sign, and both approaches were included here, referred to as CRM-CSon (where a correct response required identification of all three target words) and CRM-CSoff (where responses regarding the call sign were not required). Both approaches were assessed in terms of (1) test-retest reliability, comprising *stability* and *variability* of scores across test times (Summerfield et al, 1994); and (2) *concurrent validity*, referring to the level of agreement between scores on the CRM and scores from a similar SiN test, in this case the TDT (Kinson, 2012). *Stability* refers to the degree to which the mean SRT differs between test times. *Variability* refers to the random variation in individual SRTs between test times expressed here using the within-subject variation of a one-way analysis of variance (ANOVA) and the 95% confidence limits of this value (95% CL, Bland, 2000, p.270). *Concurrent validity* is expressed as the difference in average SRT between a CRM condition and the TDT.

# 5.1 Study 2 methods

‘Normal-hearing’ (n=30) participants (10 male, 20 female, mean age 24 years) attended two identical sessions. At each session participants completed three SiN test conditions twice: 1) CRM-CSon, 2) CRM-CSoff and 3) the TDT, presented with its own stationary speech-spectrum noise. This produced four SRTs for each condition. The order of testing was randomised between participants using a Latin Square. A two-down-one-up adaptive staircase procedure was used (Levitt, 1971). The initial SNR was –1 dB. The step size was 8 dB until the first reversal, 4 dB until third reversal and 2 dB for the remaining eight reversals. The SRT was calculated as the mean SNRs of the final eight reversals.

# 5.2 Study 2 results and discussion

Figure 2 displays box plots for the three test conditions and four test times (T1-T4). Each box plot displays the median, interquartile ranges and minimum and maximum SRTs for each test time and the results averaged across the four test times, for each condition. Table 2 shows the test-retest reliability values for the test conditions and the concurrent validity for the CRM conditions.

In terms of stability, the mean change in SRT both within and between-sessions was ≤1 dB. In terms of variability, the two conditions of the CRM were similar to each other. For clinical significance this value can be compared to the level of variability observed in two clinically used SiN which both use closed set stimuli and run in an adaptive procedure. The variability for the two CRM conditions is lower than the TDT (data collected in Study 2) and the McCormick Automated Toy Test (with a variability of 2.5 dB, reported in Summerfield et al, 1994). In addition, Smits et al (2004) calculated the variability of the Dutch TDT as the standard deviation of the individual difference across all participants and reported a value of 1.3 dB. The equivalent values calculated for the CRM conditions in the current study were ≤1 dB. In terms of concurrent validity, the SRT for CRM-CSoff condition was found to be similar to the TDT with a mean difference of 1.3 dB with a 95% CI of ± 0.3 dB. The average SRT for the CRM-CSon condition was significantly (p<.0001) higher (worse) in comparison to the TDT and the CRM-CSoff. This difference in performance is unlikely to be caused by an increase in the number of words to recall since it is widely accepted that individuals can store between four and seven simultaneous elements in their working memory at any one time (Hulme et al, 1995; Cowan, 2001). The average SRT for the call sign target word group is lower than for the colour and number, therefore reducing the average SRT for the CRM-CSon condition in comparison to the CRM-CSoff. Despite the difference in average SRT measurements, the test-retest reliability and concurrent validity of the CRM using either scoring method are very similar and are broadly comparable with similar existing tests. Kitterick et al (2010) report the SRT and between-subject variation for the CRM in a two-down-one-up adaptive procedure in which the participant responded to CRM sentences presented via a loudspeaker from a fixed location with a masking noise made up of multiple CRM phrases. The results are comparable to the CRM-CSon condition in Study 2. The between-subject variability scores (reported as between-subject standard deviations of the mean SRT) are similar to those obtained in Study 2 (1.5 dB Kitterick et al, 2010; 0.8 dB Study 2). The mean SRT scores are lower for Study 2 (–7.7 dB SNR) than the Kitterick et al (2010) study (~ –13 dB SNR, read from Figure 3 of publication), however this is expected given the different masking noises used (a fluctuating multi-talker masker will elicit lower SRTs than stationary speech-spectrum noise). Eddins and Liu (2012) reported a mean SRT of –5.7 dB SNR for the American English CRM sentences when presented in a two-down-one-up adaptive procedure in cafeteria noise, which is not significantly dissimilar to the CRM-CSon result reported in Study 2. This comparison of results provides confidence that the CRM adaptive procedure is behaving as expected in comparison to previous versions.

At this stage, Study 2 has only validated the CRM adaptive procedure with normal-hearing listeners in one type of background noise (stationary speech-spectrum) and using one type of measurement procedure (a two-down-one-up adaptive procedure). Different usages of the CRM may involve presenting the CRM stimuli in alternative background noises or using a different measurement method, impacting the average SRTs and possibly the test-retest reliability. For example the CRM stimuli could be presented in military specific background noise as part of an AFFD protocol. Giguère et al (2008) used the Hearing in Noise Test stimuli presented in occupationally relevant background noises to predict speech intelligibility for Department of Fisheries and Oceans Canada employees in a variety of communication environments. A similar approach could be adopted for the CRM, presenting the sentences in background noises relevant to the MCATs, such as a vehicle engine noise, radio noise, competing speakers or weapon systems firing (Bevis et al, 2014). Further validation of the CRM in continuous and intermittent background noises, exploring the impact of both informational and energetic maskers (Brungart, 2001b) is required before the test can be used in these contexts.

# 6. Study 3: preliminary evaluation of the adaptive CRM with hearing-impaired listeners

Our interest in using the CRM for assessing AFFD for military personnel means that it will be used for both normal-hearing and hearing-impaired participants. It cannot be assumed that the performance metrics of the test observed in Study 2 with normal-hearing participants will be the same with a hearing-impaired sample. The aim of Study 3 was to identify any major issues with the performance of the CRM with hearing-impaired military personnel using a relatively small sample ahead of a larger scale and more thorough investigation. Study 3 therefore provides a preliminary indication of the test-retest reliability and concurrent validity with hearing-impaired participants, as well as an indication that the CRM is at least broadly sensitive to hearing loss, as measured using PTA.

There is more between-subject variation in the SRTs of the hearing-impaired population in comparison to the normal-hearing samples from studies 1 and 2 (see Table 2); consequently, correlations were used in Study 3 to provide a fuller indication of the performance of the CRM. Specifically, *concurrent validity* was expressed as the correlation of SRT between the CRM and the TDT and *replicability* was calculated as the correlation between test times across participants.

# 6.1 Study 3 methods

A sample of 22 hearing-impaired military personnel (20 male, 2 female, mean age 46 years) were recruited from the Defence Audiology Service (DAS) at the Institute of Naval Medicine. Individuals were invited by post to take part if they had an audiogram completed within the last 12 months and had sensorineural hearing loss with an average threshold in the better hearing ear between > 20 and ≤ 70 dB HL. The study took place immediately after a routine audiology appointment. The spread of hearing impairment was dictated by the population attending appointments at DAS, resulting in more participants with milder hearing acuities compared to those with more severe impairments. The spread of hearing acuities grouped by better hearing ear average (see Section 3) was: 20–30 dB HL, n=9; 31–40 dB HL, n=6; 41–50 dB HL, n=5; and 51–60 dB HL, n=2.

The materials and procedures carried out in Study 3 were nearly identical to those described in Study 2, with only three differences. Firstly, only two SRT measures for each SiN test were completed because participants could only attend one session due to travel constraints; patients travel from all over the UK to visit DAS and therefore it was not feasible to ask them to attend on a second occasion within the timeframe of the study. Secondly, the noise presentation level was not fixed at 63 dB A, but varied between participants to ensure the speech was presented supra-threshold. For the TDT and CRM separately, participants listened to 10 s of continuous speech and adjusted it to a comfortable level. For each individual the relative change in volume was applied to the overall stimuli presentation level, ensuring the signal was audible. Thirdly, the initial SNR in Study 3 was set 10 dB higher in comparison to Study 2 to ensure the first presentation of the speech was intelligible.

# 6.2 Study 3 results and discussion

The distributions of the SRTs across different levels of hearing acuity are shown in Figure 3. The hearing acuity groups have been chosen for illustrative purposes only, demonstrating that as hearing acuity worsens, so do the SRT scores. As such it can be concluded that all three SiN tests are, to some extent, sensitive to hearing acuity. Due to recruitment difficulties, the hearing-impaired population did not contain an even spread of hearing acuities, in particular there were only two participants with severe hearing losses. It is therefore not possible to make assumptions about how the test performs with this population or to explore how well the versions of the CRM test are able to distinguish between different levels of hearing acuities. It is also worth noting that the difference in mean ages between the normal-hearing and hearing-impaired samples may have influenced the difference in SRT scores across the two samples (Füllgrabe et al, 2014) and a further study with aged matched samples would be desirable.

The test-retest reliability values for the hearing-impaired sample were calculated based on the assumption that test-retest reliability does not vary as a function of hearing impairment. To assess whether this is a fair assumption, Figure 4 shows the within-subject difference in scores between test times (T1 & T2) for the three tests conditions. A visual inspection of Figure 4 shows an even distribution in the difference in SRT scores between T1 and T2, providing evidence that, within this sample at least, there was no systematic effect of the magnitude of change in SRT scores across T1 and T2 with respect to SiN ability. The test-retest reliability results reported for the hearing-impaired sample are based on within-session data only because participants attended one session. It is therefore not possible to draw conclusions about the between-session test-retest reliability of the CRM conditions for this sample.

For all the test conditions, the mean change in SRT between T1 & T2 is small (<0.5 dB, see stability values in Table 2) and is not markedly worse than for the normal-hearing sample. The variability values, reported in Table 2, are also very similar to the data from the normal-hearing sample. The replicability of the data, calculated using Pearson’s correlation coefficient, is higher for the two CRM conditions than the TDT. There was a larger amount of shared variance ($r^{2}$, shown as a percentage) between test times for the CRM (CRM-CSoff $r^{2}$=76%, CRM-CSon $r^{2}$=83%) than for the TDT ($r^{2}$=41%), indicating that the replicability of the CRM is sufficient for a clinical SiN test. To summarise, for both of the CRM conditions and the TDT, the test-retest reliability is not significantly worse for the hearing-impaired sample in comparison to the normal-hearing participant data. In addition, the test-retest reliability of the CRM is similar, and for some performance metrics marginally better, than the TDT. This supports the conclusion that the CRM test conditions produce robust SRT measurements.

The concurrent validity between the CRM conditions and the TDT was assessed by calculating Pearson’s correlation coefficient between the mean scores on the two tests. For both of the CRM test conditions the effect sizes can be described as ‘large’ ($r $≥ 0.5) (Cohen, 1988), indicating that around half of the variance ($r^{2}$, shown as a percentage) observed in the CRM test scores is shared with that observed in the TDT scores (CRM-CSoff $r^{2}$=43%, CRM-CSon $r^{2}$=52%). In addition, the difference between the SRT scores on the CRM conditions and the TDT are fairly small, further supporting the conclusion that the CRM test conditions are in fact assessing SiN ability.

The results from studies 2 and 3 have not shown either scoring methods to display significantly better test-retest reliability or concurrent validity (Table 2). The scoring methods are highly correlated (the correlation coefficient between the first test time for CRM-CSon and CRM-CSoff is 0.88, 95% upper and lower confidence limits of 0.72 – 0.95). It is concluded that running the adaptive procedure with either scoring method is suitable for use as a method for predicting SRTs, extending the options for how the CRM test can be used.

Although it has been established that the CRM is a robust measurement tool it can only be deemed a useful measure of AFFD if it is able to accurately discriminate between individuals who have sufficient hearing to carry out the MCATs to a predefined satisfactory level. In order to explore this further it is proposed that simulations of the speech-communication MCATs are developed with the ultimate goal of using these simulations to measure and compare the predictive validity of the CRM and PTA as measures of AFFD.

It is worth noting that the relatively small sample size used in Study 3 increases the risk that the within-subject variation between test times may have been underestimated and therefore not be representative of the wider hearing-impaired military population. Furthermore, in order to fully understand the distribution of SRTs across different levels of hearing acuity further work is required with a larger sample of hearing-impaired listeners.

# 7. Future work: evaluating the CRM as a tool for assessing AFFD

The primary aim of this paper was the development and evaluation of the new British English version of the adaptive CRM with stationary speech-spectrum noise. The key motivation for this work was the development of a SiN test as a potential tool for predicting AFFD for military personnel. In this paper the CRM has been shown to accurately measure SRTs and can be regarded as a ‘ready to use’ SiN test, however further work is required to explore the CRM as a tool for measuring AFFD.

In order to evaluate the CRM as a novel hearing-screening measure to be used for occupational AFFD assessment it is necessary to investigate the relationship between performance on the CRM and performance when carrying out ‘real world’ occupational listening tasks, such as those described by the MCATs (Semeraro et al, 2015). Further work is required to determine a method for assessing ‘real world performance’ on these tasks to enable evidence based AFFD standards to be generated.

There are several benefits of introducing a SiN test as a hearing conservation assessment tool, not just within the Armed Forces, but also in other noisy occupations where regular hearing assessment is important. SiN testing can be carried out in an environment with higher ambient noise levels than allowed during PTA, increasing the flexibility of the test environment and therefore the feasibility of regular assessment within the work place. The CRM adaptive procedure is fully automated, removing the need for a specially trained tester, allowing for multiple individual’s to be simultaneously tested. One of the key reasons for selecting the CRM as alternative AFFD measure to PTA was the face validity of the test.

# 8. Summary and conclusions

It has been demonstrated that the British English adaptive CRM using stationary speech-spectrum noise is a ‘ready to use’ SiN test, displaying good test-retest reliability and concurrent validity (when compared to the TDT), with two scoring methods (call sign on and call sign off) and it is sensitive to hearing impairment.

In conclusion:

* The CRM test can be used to accurately measure individual SRTs.
* Due to better face validity and the potential to measure supra-threshold hearing deficits, rather than just audibility, the CRM should be investigated as a measure of AFFD for military personnel.

*The CRM will be publically available as a tablet application to be run on an Apple iPad for research purposes only. The audio files (wav format) can also been accessed through the IJA website. Details about the release date, how to download the application and data from studies one and two can be found at* [*www.southampton.ac.uk/engineering/crm*](http://www.southampton.ac.uk/engineering/crm)*.*

# 9. Acknowledgements

Thank you to the Surgeon General of the Ministry of Defence for funding this work and to the military personnel and civilian participants.

# 10. References

Akeroyd, M.A., 2008. Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *Int J Audiol*, 47(2), S53–S71.

Anastasi, A., 1988. Psychological Testing. 6th ed., New York: Macmillan.

Bench, J., Kowal, A. & Bamford, J., 1979. The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *British Journal of Audiology*, 13(3), 108–12.

Bevis, Z.B., Semeraro, H.S., Rowan, D., van Besouw, R., Allsopp, A., 2014. Fit for the frontline? A focus group exploration of auditory tasks carried out by infantry and combat support personnel. *Noise Health*, 16(69), 127–135.

Bland, M., 2000. An introduction to medical statistics 3rd ed., Oxford: Oxford University Press.

Bolia, R.S., Nelson, W.T., Ericson, M. a, Simpson, B.D., 2000. A speech corpus for multitalker communications research. *J Acoust Soc Am*, 107(2), 1065–6.

Brungart, D.S., 2001a. Evaluation of speech intelligibility with the coordinate response measure. *J Acoust Soc Am*, 109(5), 2276-2279.

Brungart, D.S., 2001b. Informational and energetic masking effects in the perception of two simultaneous talkers. *J Acoust Soc Am*, 110(5), 2527–2538.

Brungart, D.S., Sheffield, B., 2013. Development of Auditory Fitness for Duty Standard for Military Operations. In: NATO Science and Technology Organisation, *Human Factors and Medicine Panel Symposium*. Milan, Italy, 15-17 April 2013.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences 2nd ed., Hillsdale, New Jersey: Lawrence Erlbaum Associates Inc Publishers.

Cowan, N., 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behav Brain Sci*, 24(1), 87–114.

Eddins, D. A., Liu, C., 2012. Psychometric properties of the coordinate response measure corpus with various types of background interference. *J Acoust Soc Am*, 131(2), 177–183.

Füllgrabe, C., Moore, B, C.J., Stone, M.A., 2014. Age-group differences in speech identification despite matched audiometrically normal hearing: contributions from auditory temporal processing and cognition. *Front Aging Neurosci*, 6(347), 1–25.

Giguère, C., Laroche, C., Soli, S.D., Vaillancourt, V., 2008. Functionally-based screening criteria for hearing-critical jobs based on the Hearing in Noise Test. *Int J Audiol*, 47(6), 319–28.

Hulme, C., Roodenrys, S., Brown, G., Mercer, R., 1995. The role of long-term memory mechanisms in memory span. *Br J Psychol*, 86(4), 527–536.

Kingdom, F., Prins, N., 2010. Psychophysics: A Practical Introduction 1st ed., London: Elsevier Limited.

Kinson, T., 2012. *The tech behind the check, Action on Hearing Loss Blog*. [online] Available at: <http://www.actiononhearingloss.org.uk/community/blogs/our-guest-blog/the-tech-behind-the-check.aspx> [Accessed 8 June 2016].

Kitterick, P.T., Bailey, P.J., Summerfield, A.Q., 2010. Benefits of knowing who, where, and when in multi-talker listening. *J Acoust Soc Am*, 127(4), 2498–2508.

Kujawa, S.G., Liberman, M.C., 2015. Synaptopathy in the noise-exposed and aging cochlea: Primary neural degeneration in acquired sensorineural hearing loss. *Hear Res*, 330, 191–199.

Leek, M.R., 2001. Adaptive procedures in psychophysical research. *Percept Psychophys*, 63(8), 1279–1292.

Leensen, M.C.J., 2013. *Noise induced hearing loss: Screening with pure-tone audiometry and speech-in-noise testing.* Ph.D. University of Amsterdam. Available at: <<http://dare.uva.nl/record/1/394362>> [Accessed 8 June 2016].

Leensen, M.C.J., de Laat, J. a P.M., Dreschler, W. a, 2011a. Speech-in-noise screening tests by internet, part 1: test evaluation for noise-induced hearing loss identification. *Int J Audiol*, 50(11), 823–834.

Leensen, M.C.J., de Laat, J.A.P.M., Snik, A.F.M., Dreschler, W.A., 2011b. Speech-in-noise screening tests by internet, Part 2: Improving test sensitivity for noise-induced hearing loss. *Int J Audiol*, 50, 835–848.

Levitt, H., 1971. Transformed Up-Down Methods in Psychoacoustics. *J Acoust Soceity Am*, 49(2), 467-477.

Lutman, M.E., Hall, S.J. & Athalye, S., 2006. Development of a telephone hearing test. *Proceedings of the Institute of Acoustics,* 28(1), 240–43.

Nilsson, M., Soli, S.D. & Sullivan, J.A., 1994. Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *J Acoust Soceity Am,* 95(2), 1081–99.

Ozimek, E., Kutzner, D., Sęk, A., Wicher, A. 2009. Development and evaluation of the Polish Triplet Test. *Speech Communication,* 51(4), 307-316.

Peters, R.W., Moore, B.C.J., Baer, T., 1998. Speech reception threshold in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. *J Acoust Soc Am*, 103(1), 577–587.

Plack, C.J., Barker, D., Prendergast, G., 2014. Perceptual Consequences of “Hidden” Hearing Loss. *Trends Hear*, 18, 1–11.

Plomp, R., 1978. Auditory handicap of hearing impairment and the limited benefit of hearing aids. *J Acoust Soc Am*, 63, 533–549.

Semeraro, H.D., Bevis, Z.L., Rowan, D., van Besouw, R.M., Allsopp, A.J., 2015. Fit for the frontline? Identification of mission-critical auditory tasks (MCATs) carried out by infantry and combat-support personnel. *Noise Health*, 17(75), 98–107.

Shinn-Cunningham, B., Bharadwaj, H., Choi, I., Goldberg, H., Masud, S., et al, 2014. Quantifying supra-threshold sensory deficits in listeners with normal hearing thresholds. *J Acoust Soc Am*, 136(4), 2258–2258.

Smits, C., Kapteyn, T.S., Houtgast, T., 2004. Development and validation of an automatic speech-in-noise screening test by telephone. *Int J Audiol*, 43, 15–28.

Smoorenburg, G.F., 1992. Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram. *J Acoust Soc Am*, 91(1), 421–437.

Summerfield, Q., Palmer, A.R., Foster, J.R., Marshall, D.H., Twomey, R., 1994. Clinical evaluation and test-retest reliability of the IHR-McCormick Automated Toy Discrimination Test. *Br J Audiol*, 28, 165–179.

Tufts, J.B., Vasil, K. A., Briggs, S., 2009. Auditory Fitness for Duty: A Review. *J Am Acad Audiol*, 20(9), 539–557.

Wilson, R.H. & Cates, W.B., 2008. A comparison of two word-recognition tasks in multitalker babble: speech Recognition in Noise Test (SPRINT) and Words-in-Noise Test (WIN). *J Am Acad Audiol*,19(7), 548–56.

World Health Organisation (WHO), 2016. *Screening for various cancers* [online] Available at: http://who.int/cancer/detection/variouscancer/en/ [Accessed May 27, 2016].

# 11. Figure captions

***Figure 1*** *Psychometric functions for individual target words with the RMS amplitude adjusted recordings from Sessions 1 and 2 (left) and the SRT50 matched recordings from Session 3 (right).*

***Figure 2*** *Box plots showing normal-hearing SRTs (for each test time and averaged across all four test times, labelled T1-4) for both CRM conditions and the TDT (n=30). Outliers exceed 1.5 times the interquartile range****.***

***Figure 3*** *Box plots showing the distribution of the SRTs for different levels of hearing acuity for the two CRM conditions methods and the TDT. Participants’ results were grouped by hearing acuity based on their averaged pure-tone thresholds: normal ≤20 dB HL, n=30 (civilians from Study 2); mild >20 and ≤35 dB HL, n=12; moderate to severe >35 and ≤55 dB HL, n= 10.*

***Figure 4*** *Difference in SRT scores between test time one (T1) and test time two (T2) across the hearing-impaired sample for each SiN test, to display changes in stability across the sample. One anomaly has been removed from the figure for ease of plotting (TDT, x=0.75, y=* –*9).*

# 12. Tables

***Table 1*** *The British-English CRM speech corpus used in the current study*

|  |  |  |
| --- | --- | --- |
| ***“*Ready *call sign*** | **go to *colour*** | ***number* now*”*** |
| Alpha | Oscar | Black | One |
| Bravo | Papa | Blue | Two |
| Charlie | Quebec | Brown | Three |
| Delta | Tango | Gold | Four |
| Echo | Victor | Green | Five |
| Foxtrot | Whisky | Grey | Six |
| Hotel | X-Ray | Pink | Eight |
| Kilo | Yankee | Red | Nine |
| Lima | Zulu | White | Ten |

***Table 2*** *Test-retest reliability (stability, variability and replicability) and the concurrent validity of the CRM (CSon and CSoff) and TDT for normal-hearing (NH, four test times, T1-T4) and hearing-impaired (HI, two test times, T1 and T2) participants.*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **CRM-CSoff** | **CRM-CSon** | **TDT** |
| **Normal-hearing****(Study 2)** | **Test-retest reliability** | **Stability1 (dB)** | *T1/T2* | *T3/T4* | *T1/T3* | *T1/T2* | *T3/T4* | *T1/T3* | *T1/T2* | *T3/T4* | *T1/T3* |
| 0.3(± 0.5) | –0.1(± 0.4) | 0.7(± 0.5) | 0.6(± 0.5) | 0.0(± 0.5) | 1.0(± 0.6) | 0.4(± 0.6) | 0.2(± 0.6) | 0.9(± 0.6) |
| **Variability2 (dB)** | 1.9 | 2.1 | 2.6 |
| **Concurrent validity3 (dB)** | 1.3 (± 0.3) | 4.0 (± 0.3) | NA |
| **Hearing-impaired****(Study 3)** | **Test-retest reliability** | **Stability4 (dB)** | 0.4 (± 0.8) | 0.3 (± 0.6) | –0.5 (±1.3) |
| **Variability5 (dB)** | 2.5 | 1.7 | 3.9 |
| **Replicability6 (*r*)** | *r =* 0.87 (0.71–0.94)*r2*=0.76 | *r =* 0.91(0.79–0.96)*r2*=0.83 | *r =* 0.64 (0.30–0.84)*r2*=0.41 |
| **Concurrent validity** | **7a (dB)** | –0.5 (± 1.0) | 2.3 (± 0.9) | NA |
| **7b (*r & r2*)** | *r* = 0.65 (0.31 – 0.84)*r2*=0.42  | *r* = 0.72 (0.43 – 0.88)*r2*=0.52 | NA |
| ***Table Notes*** |
| **Normal-hearing****(n=30)** | 1. *Stability: mean change in SRT between test times and the 95% confidence interval of the mean for Session 1 (T1 minus T2), Session 2 (T3 minus T4) and between sessions (T1 minus T3).*
2. *Variability: the 95% confidence interval of the true SRT score any one measurement value (Bland, 2000, p269), measured across T1-T4.*
3. *Concurrent validity: difference (in dB) between SRT scores (mean of T1-T4) for the CRM conditions and TDT. A positive number indicates a lower SRT for the TDT in comparison to the CRM test condition and a negative number indicates a higher SRT for the TDT in comparison to the CRM test condition.*
 |
| **Hearing-impaired****(n=22)** | 1. *Stability: mean change in SRT between T1 and T2 and the 95% confidence interval of the mean*
2. *Variability: the 95% confidence interval of the true SRT score any one measurement value (Bland, 2000, p269), measured across T1 and T2.*
3. *Repeatability: Pearson’s correlation coefficient between T1 and T2 SRT scores*

***7a.*** *Concurrent validity: difference (in dB) between SRT scores (mean of T1 and T2) for the CRM conditions and TDT (see point 3).****7b.*** *Concurrent validity: Pearson’s correlation coefficient**between the CRM test condition SRT scores (mean of T1 and T2) and the TDT.*  |